Author's response to reviews

Title: PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers

Authors:

Roy RL Bastien (roy.bastien@hci.utah.edu)
Álvaro Rodríguez-Lescure (alescure@terra.es)
Mark TW Ebbert (mark.ebbert@hci.utah.edu)
Aleix Prat (aleixprat@yahoo.es)
Blanka Munárriz (munarriz_bla@gva.es)
Leslie Rowe (leslie.rowe@aruplab.com)
Patricia Miller (patty.miller@aruplab.com)
Manual Ruiz-Borrego (manuel.ruiz.borrego.sspa@juntadeandalucia.es)
Daniel Anderson (daniel.anderson@aruplab.com)
Bradley Lyons (lyonsbw@aruplab.com)
Isabel Álvarez (isabelmanuela.alvarez@osakidetza.net)
Tracy Dowell (dowellt@aruplab.com)
David Wall (wallde@aruplab.com)
Miguel A Segui (mseguui@tauli.cat)
Lee Barley (barleyl@aruplab.com)
Kenneth M Boucher (kenneth.boucher@hci.utah.edu)
Emilio Alba (ealbac@uma.es)
Lisa Pappas (lisa.pappas@hci.utah.edu)
Carole A Davis (carole.davis@hci.utah.edu)
Ignacio Aranda (aranda_ign@gva.es)
Christiane Fauron (christiane.fauron@aruplab.com)
Inge J Stijleman (inge.stijleman@hci.utah.edu)
José Palacios (jose.palacios.sspa@juntadeandalucia.es)
Antonio Antón (aantont@salud.aragon.es)
Eva Carrasco (evacarrasco@geicam.org)
Rosalía Caballero (rcaballero@geicam.org)
Matthew J Ellis (mellis@dom.wustl.edu)
Torsten O Nielsen (torsten@mail.ubc.ca)
Charles M Perou (chuck_perou@med.unc.edu)
Mark Astill (astillme@aruplab.com)
Philip S Bernard (philip.bernard@hci.utah.edu)
Miguel Martin (mmartin@geicam.org)

Version: 2 Date: 23 June 2012

Author's response to reviews: see over
Dear Editor,

The authors appreciate the thoughtful review of our manuscript titled “PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers”. We have addressed each of the reviewer’s concerns and are pleased to submit this revised and improved manuscript.

Reviewer #1: Thomas Hughes
Overall, I find the study to be potentially worthwhile, although I have some doubts about whether there are any novel scientific findings beyond the validation of the new assay itself. However, I find considerable problems with the article including a lack of detail on key aspects of what was done and why, lack of focus on the specific findings that the authors wish to highlight, and some over-statements or even exaggerations.

Response: The authors have made significant edits to improve the focus of the manuscript, emphasize findings from the large 814 sample test set (rather than the training set validation), and have removed statements about PAM50 being a standard since it is not widely used in Europe.

Major points
1) Abstract: A key finding seems to be “RT-qPCR scoring… provided more prognostic information than histological-molecular scoring”. I think this claim relies solely on Figure 6. I find this figure to be rather difficult to interpret since hormone receptor status is not broken down into ER and PR separately and because of the confounding factor that the cohort includes patients who received a variety of therapies (the two different arms of the chemotherapy trial; also with or without hormonal therapies – a fact not mentioned in this manuscript anywhere). The hormone therapy is particularly problematic since this was presumably given on the basis of ER expression as detected by IHC – and therefore is specific for some groups shown on the right, but is probably variable in all groups on the left. Could the authors justify this conclusion more thoroughly in the relevant results section?

Response: The authors appreciate this insightful comment. In Methods (Samples and Clinical Data), we now mention that patients scored ER+ by IHC did indeed receive tamoxifen. In Results, (Prognostic Significance of Gene Expression versus Standard Methods for ER and Her2 Status), we have stratified by ER and Her2 status rather than hormone status and have added an additional Figure (6C) which shows outcome based on concordant and discordant calls between the methods. Furthermore, we have added a paragraph to the Discussion stating that “despite the fact that patients were treated in favor of the IHC diagnosis (i.e. ER+ disease was treated with adjuvant tamoxifen) the course of disease was in agreement with the gene expression determination.”.
2) Throughout the manuscript there is a difficulty in articulating how the accuracy of one method of determining cancer subtypes is being assessed. It is important to note that when assigning a tumour to a subtype there is not necessarily a “right answer”. All that can be done is one method can be compared to another method – but it is not really possible to say which is right. For example, in the abstract, “IHC/CISH biomarkers were not able to accurately identify all intrinsic subtypes”. It is true that the results from IHC differ from those of qPCR – but it is not fair to say that the qPCR must be correct and IHC has failed. This difficulty is embedded within the writing – especially within the “Subtype, immunohistochemistry and RT-qPCR…” section of the results.

Response: The authors agree that the terminology for “accuracy” could be confusing since it can be used to refer to agreement between methods or having the correct answer in terms of outcome. We have removed the use of the word “accurate” and use concordance or agreement instead when comparing methods of biomarker assessment.

3) Introduction, 2nd para. It is simply not true that the PAM50 gene set has become the gold standard for subtyping (at least certainly not in the UK, and I doubt in many other countries). Nor is it “the standard method of classification” as claimed in the Discussion.

Response: The authors have removed referring to the PAM50 assay as a standard throughout the manuscript.

4) Methods, top page 9. How were IHC scores dichotomised into positive and negative?

Response: We have clarified throughout the manuscript that the ROC curve figures were generated based on the IHC scores (positive vs negative) and the continuous qRT-PCR data. However, the cut-points shown on the ROC curves for the test set were those fixed from the training set, which were selected from expression across the subtypes.

5) I don’t feel the NKI analyses, which are barely mentioned in the results text, contribute enough to justify their inclusion.

Response: We have removed the NKI figure and results discussing concordance between the microarray and RT-qPCR classifications.

6) Results, “Training set” and Figure 1. Could the authors more fully explain why non-neoplastic tissue tissue should be included in a training set for analysis of tumours? Could the authors please highlight the 16 non-neoplastic samples on Figure 1 – are these within the right hand group of 17 samples, therefore presumably one tumour has grouped with the non-neoplasics? Or are they elsewhere? In the text the green highlighted samples are described as “normal-like” (a tumour classification) whereas I think they are probably actually non-neoplasics (as suggested by the figure legend). The phrase “The accuracy of the RT-qPCR training set in correctly
assigning tumor subtype was 93% as determined by comparing to our previously reported research training set that used microarray data” is an example of my point 2 – surely the authors simply mean concordance, not accuracy or correctly. I’m not completely sure of the relevance of the statement anyway as the data are not shown.

Response: Justification for using non-neoplastic breast tissue to represent a “Normal” subtype is more clearly presented throughout the manuscript. The dendrogram for Figure 1 has been enlarged to more clearly identify the group. We have removed the use of the word accuracy as mentioned previously and have removed the concordance between the RT-qPCR subtype classification and microarray.

7) Results “Interference from normal...”. Which non-neoplastic sample was used (ie which on Figure 1) and did it matter which with respect to the change in classification seen? Which tumour subtype RNA were used? “The switch from Luminal B to Luminal A required less than 50% contribution from the normal breast tissue signature”, this is directly contradicted in the Discussion. A very surprising result is seen in Additional file 6 – where the esr1 call changes from intermediate to low and back as additional non-neoplastic RNA is added to the HER2-E sample; please would the authors comment.

Response: It has now been clarified in Results what samples were used for prototype groups in the Interference section. The Results and Discussion now state that switches were seen at 50%. Furthermore, we provide an explanation for the switching in the ESR1 Score as non-neoplastic RNA is added to the HER2-E subtype.

8) Results “Validation of single and...”. How was it decided that medians were appropriate cut offs for some markers and quartiles for others? To what is high/intermediate/low in red/yellow/blue for proliferation and positive/negative in red/blue for “luminal” referring in additional file 7? Could the authors explain why these plots “validate” the cut offs (as in the title of this section), rather than merely demonstrating their effect?

Response: The different cut-points were selected since they approximated IHC status in the subtypes. For instance, rarely was a Luminal subtype low ER. The proliferation cut-points were solely based on quartile distribution across all subtypes. Reference to the luminal scores has been removed since it provided little utility beyond the ESR1 score. We have tried to clarify that cut-points from the training set provide good sensitivity/specificity when applied to an independent test set; thus, we can apply these cut-points across different cohorts without having to re-optimize.

9) Results “Subtype, immunohistochemistry...” and Figure 3. Could the authors explain more fully and clearly what the ROC curves demonstrate. They say that “Intermediate and high gene scores were combined to dichotomise the continuous gene expression data” – giving the impression that they have used the lower cut offs in Table 2. But I
believe they have used the ROC curves to test different cut offs for the qPCR data basing these sensitivity/specificity analyses on the positives defined by IHC/ISH. These analyses would certainly be appropriate – but I've had to guess that this is what was done. Most worryingly, the graphs seem to demonstrate that the best cut offs have not been selected – particularly in graph C, where the cut off has inflated the false positive rate for no gain in true positive rate. The rest of this section is very difficult to read for at least two reasons. A - it is consistently unclear exactly what the authors mean by each description of a patient group – for example, “Luminal A tumours more frequently...”. Do the authors mean luminal A as defined by their PAM50 classification – or by IHC staining? This needs to be very clear since different groupings are defined by the two technologies and since the findings here are counter-intuitive with some tumours found to be negative for the classic markers usually associated with their group. B – there is simply too much detail and too few take home messages. What are the main important points the authors wish to make?

Response: We have re-written this Results section to make it clear that the IHC data is split into clinically positive and negative and compared to the continuous RT-qPCR expression. The area under the curve illustrates agreement between the methods. The actual cut-points selected come from the quartile distributions across the subtypes. We always refer to subtypes (i.e. Luminal A) as expression-based subtypes or otherwise refer to IHC panels. We have tried to make this section of the Results more succinct.

10) Discussion, page 17. “We used a novel approach for selecting... cut offs”. I am not clear that using either median or quartile cut offs is novel or was “shown to be reproducible in an independent data set”. Could the authors explain this more fully?

Response: The novelty was not in using median or quartile cut-points but rather using the distribution of expression across the subtypes as the basis for the cut-points. Nevertheless, we have removed the word novel from the sentence. We found that the distribution of expression across the subtypes was similar between the training and test sets allowing us to apply fixed cut-offs.

11) Discussion, page 18. I believe the discussion of transcription factors should be omitted since these data are not a focus of the manuscript.

Response: We have removed the table and discussion about other genes important for distinguishing the subtypes.

12) Conclusions. The surrogate subtyping is barely mentioned in the results text yet is highlighted as a key conclusion – please could the authors either reconsider this conclusion or more fully explain the data. The final sentence is without justification and should be removed.
Response: We discuss the discordance between single marker panels and PAM50 gene expression subtypes in order to highlight that they should not be used synonymously.

Minor points
1) The abstract is filled with jargon and abbreviations that are likely to be uninterpretable to many readers. For example: what is CLIA, what does “central scoring” mean, what is IHC/CISH?

Response: Terminology has been clarified.

2) Methods, page 10. “due to a lower concentration of a single gene within a sample”. Please change to gene product, or transcript or something similar.

Response: This phrase has been removed.

3) Please increase the size of the dendrogram relative to the heat map in Figure 1.

Response: Dendrogram has been increased in size.
Reviewer #2: Christopher Wong

1. As several academic papers on PAM50 have been previously published, I was expecting to see data demonstrating the utility of this signature for clinical use on individual patients, especially since the authors emphasize that this RT-qPCR was developed in a CLIA laboratory. Moreover, Nanostring announced in Dec 2011, the largest clinical validation of the PAM50 samples (on 1000 patients from the TransATAC study) [http://www.nanostring.com/file/press/press_items_link_url_89.pdf].

Response: The PAM50 was initially developed on the RT-qPCR platform and transitioned to the clinical lab for single site testing. The gene set has also been licensed to Nanostring to be de-centralized and widely-distributed as an IVD. The press release by Nanostring on the TransATAC study was the beginning of their clinical trial work, while many more trials have already been completed using the RT-qPCR format (MA.5, MA.12, GEICAM/9906). The emphasis of this manuscript is to use the large GEICAM/9906 study that had centralized IHC scoring to compare IHC panels to PAM50 subtyping and single gene expression scores.

2. This paper demonstrates the utility of the PAM50 signature for profiling a set of 814 patient samples from a phase III clinical trial. However it is not too different from the previous papers describing the PAM50 methodology, particularly Nielsen et al, 2010.

Response: The PAM50 RT-qPCR methodology was same between the Nielsen et al manuscript and GEICAM/9906; however, the Nielsen manuscript was emphasizing subtyping in an ER+ only cohort given tamoxifen and no chemotherapy. The GEICAM/9906 cohort is locally advanced primary invasive breast cancers (ER+ and ER-) that were randomized to an anthracyline-based regimen (AC = doxorubicin-cyclophosphamide) with or without a taxane (paclitaxel). Patients scored ER+ by IHC were given tamoxifen. This study provided an excellent test set for comparing IHC status to single gene expression scores show that despite the fact that patients were treated in favor of the IHC diagnosis (i.e. ER+ disease was treated with adjuvant tamoxifen) the course of disease was in agreement with the gene expression determination. This has been made clear in the Results and Discussion.

3. It is not apparent if the subtype scoring system used in this paper is the same as the ROR score method previously published (Cheang et al, JNCI 2009) or if they have developed a new method which would integrate the results from all 50 genes into a classification/predictor system that would work for a single patient.

Response: The ROR Score is not a component of this study since it was derived from microarray data and would need to be re-derived on a different untreated patient cohort using RT-qPCR. The main emphasis of this study is comparing IHC to single gene measurements and to subtypes. We find that there is significant disagreement between PAM50 subtype assignments and using a combination of single gene markers or IHC panels to infer subtype.
4. According to description in this paper, the analysis method for subtype classification requires the use of hierarchical clustering of a population of different patient samples. For example, looking at Table 3 and Table 4, it is not apparent to me how one can figure out the subtype of the tumor by looking at the single gene scores for ESR1, PGR, ERBB2 or metagene score for proliferation, for a given patient. Typically in the clinic, one would not be testing hundreds of breast cancer samples at a time.

Response: Subtype assignments on the GEICAM/9906 test set were done without Clustering. As described in Methods the training set is initially Clustered for defining prototype samples for each subtype. After the centroids are developed in the training set, we perform single sample predictions on the test set. This has been clarified in Methods.

5. While the authors submitted many data files and tables, they did not actually submit any data to demonstrate the performance of the PAM50 RT-PCR assay, as compared to established gold standards. For eg in the last Results paragraph, authors state that “there was high agreement between IHC/CISH and RT=qPCR measurements for ER, PR, and HER2”, but did not present the actual concordance/discordance rates. What was the concordance between this PAM50 subtype classification assay for the 814 patient samples with current gold standard? Was it better or worse than the 93% accuracy for the training set (which compared to microarray data using the same gene set, rather than pathology)?

Response: The concordance between RT-qPCR gene scores and IHC status is now extensively discussed in the Results. Comparison of the RT-qPCR subtype assignment to the NKI microarray dataset has been removed based comments from another reviewer.

6. The authors, in additional file 3, set clear pass/fail criteria for the assay, but in the paper did not present how often the runs or assays had to be repeated. Eg. out of the 814 patient samples, were they able to obtain data from all these samples on the first try, or repeats were necessary, and were there any samples which could not be processed?

Response: The overall success rate from sample extraction to successful RT-qPCR was similar to previously reported studies. Overall, approximately 85-90% pass performance criteria the first time.

7. The LOD and PCR efficiency work (additional file 3) was comprehensive. However, it would be interesting to address at least in the discussion section the differences between this RT-PCR assay, and the Nanostring technology, which appears to not require PCR amplification.
Response: The authors now provide technical differences between the Nanostring nCounter and RT-qPCR platforms in the Discussion.